Detection of Traumatic Pediatric Elbow Joint Effusion Using a Deep Convolutional Neural Network

Joseph R. England[1]
Jordan S. Gross[1]
Eric A. White[1]
Dakshesh B. Patel[1]
Jasmin T. England[2]
Phillip M. Cheng[1]

OBJECTIVE. The purpose of this study is to determine whether a deep convolutional neural network (DCNN) trained on a dataset of limited size can accurately diagnose traumatic pediatric elbow effusion on lateral radiographs.

MATERIALS AND METHODS. A total of 901 lateral elbow radiographs from 882 pediatric patients who presented to the emergency department with upper extremity trauma were divided into a training set (657 images), a validation set (115 images), and an independent test set (129 images). The training set was used to train DCNNs of varying depth, architecture, and parameter initialization, some trained from randomly initialized parameter weights and others trained using parameter weights derived from pretraining on an ImageNet dataset. Hyperparameters were optimized using the validation set, and the DCNN with the highest ROC AUC on the validation set was selected for further performance testing on the test set.

RESULTS. The final trained DCNN model had an ROC AUC of 0.985 (95% CI, 0.966–1.000) on the validation set and 0.943 (95% CI, 0.884–1.000) on the test set. On the test set, sensitivity was 0.909 (95% CI, 0.788–1.000), specificity was 0.906 (95% CI, 0.844–0.958), and accuracy was 0.907 (95% CI, 0.843–0.951).

CONCLUSION. Accurate diagnosis of traumatic pediatric elbow joint effusion can be achieved using a DCNN.

Upper extremity trauma is a common chief complaint for pediatric patients presenting to the emergency department. Among these patients, acute elbow fracture is an important consideration, and elbow radiographs are the examination of choice to exclude this diagnosis. Acute pediatric elbow fracture is often a challenging diagnosis because the elbow contains multiple cartilaginous ossification centers, which are radiolucent in infancy and ossify at different rates, leading to a highly variable appearance, even for normal elbows [1]. Nondisplaced fractures can be exceedingly difficult or impossible to detect, and, in many cases, the only sign of acute elbow fracture is the presence of elbow effusion. Elbow effusion is detectable on lateral radiograph by the secondary elevation of periarticular fat by synovial fluid in the olecranon, coronoid, and radial fossae [2]. Elevation of fat in the olecranon fossa, the posterior fat pad sign, was first reported in 1954 as a sign of elbow effusion and occult elbow fracture [3]. Fat elevation in the superimposed coronoid and radial fossae, the anterior fat pad sign, was described shortly thereafter [4]. The reported frequency of occult or initially missed acute fracture in pediatric patients with traumatic elbow effusion has ranged from 17% to 77% [5–7]. Because of this high frequency, patients with isolated traumatic elbow effusion at initial radiography are often treated presumptively with elbow immobilization and follow-up radiography in 7–14 days [1, 7].

The conspicuity of elbow effusion is highly dependent on the volume of fluid and the positioning of the elbow on lateral radiographs, and radiographic diagnosis of elbow joint effusion can be especially challenging for nonradiologists [2, 8, 9]. Given that pediatric elbow radiographs are often initially interpreted by nonradiologist acute care professionals during off-hour times, accurate automated diagnosis using a computer vision algorithm may be of practical value. Potential uses for such an algorithm include decision support for radiologists in training or nonradiologists practicing in the acute setting and intelligent management of the radiology worklist, such as flagging of potential-

Keywords: artificial intelligence, deep convolutional neural network, deep learning, elbow effusion, machine learning

doi.org/10.2214/AJR.18.19974

Received April 9, 2018; accepted after revision May 17, 2018.

[1] Department of Radiology, Keck School of Medicine of USC, 1441 Eastlake Ave, Ste 2315B, Los Angeles, CA 90033. Address correspondence to P. M. Cheng (Phillip.Cheng@med.usc.edu).

[2] Department of Emergency Medicine, Harbor-UCLA Medical Center, Torrance, CA.

AJR 2018; 211:1361–1368

0361–803X/18/2116–1361

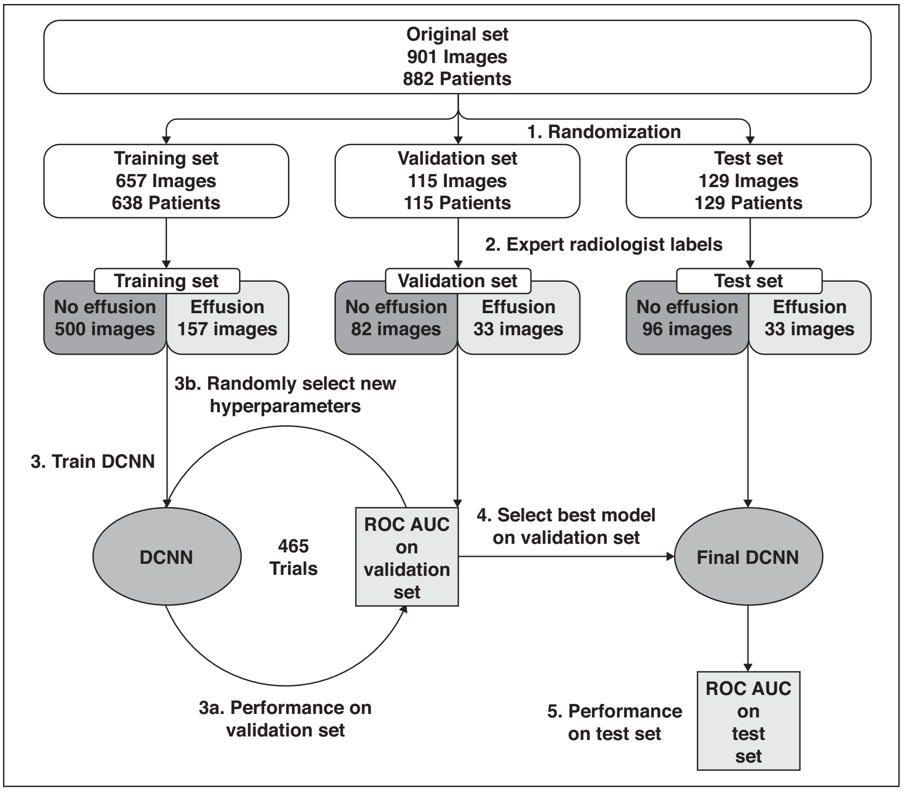

Fig. 1—Flow diagram of study method. DCNN = deep convolutional neural network.

ly positive cases for expedited interpretation by a fully trained radiologist.

Recent developments in the field of machine learning make such an algorithm possible. Artificial neural networks are a type of machine learning algorithm composed of multiple layers of computations, inspired by the architecture of biologic neural networks [10]. Convolutional neural networks are a type of artificial neural network designed to process spatial data such as images [11]. Convolutional neural networks learn directly from images by performing hierarchic feature extraction; earlier layers extract simple features, such as edges or textures, directly from the pixel data, and later layers combine the earlier simple features into increasingly complex features important for image classification [12]. Breakthrough performances in image classification by convolutional networks of increasing depth (i.e., increasing number of layers) has led to a growing branch of machine learning called deep learning [13, 14]. In medical imaging, deep learning has produced algorithms capable of expert or near-expert level performance in some medically relevant tasks. Examples include automated diagnosis of hip fracture and pulmonary tuberculosis and, more broadly, classification of chest radiographs and musculoskeletal radiographs as normal or abnormal [15–18].

The aim of this study is to explore the feasibility of using deep learning for the automated diagnosis of pediatric traumatic elbow effusion by training deep convolutional neural network (DCNN) models on a training dataset of lateral elbow radiographs and comparing the final performance on an independent test dataset to the majority opinion of three subspecialty-trained musculoskeletal radiologists.

Materials and Methods

This study was performed in accordance with the Declaration of Helsinki and was approved by the institutional review board of the University of Southern California Health Sciences Campus. Adult participant consent was not required, and parent, guardian, or next of kin consent was not required for minors because the research involved no more than minimal risk to the subjects, and the research could not practicably have been performed without the waiver of consent.

The following sections describe the process of dataset collection and labeling, the method of training the DCNN models, and the statistical analysis of model performance. A summary of these methods is depicted in Figure 1.

Data Collection and Labeling

Data for this retrospective study were obtained from our PACS. A search was performed to identify all elbow radiographs performed on pediatric patients, between 12 months and 19 years old, who were imaged and subsequently discharged from the emergency department between January 1, 2014, and September 25, 2017. Images and associated reports were examined consecutively by a postgraduate year (PGY) 5 radiology resident and were included in the dataset if the indication was blunt upper extremity trauma, the lateral radiograph was technically adequate, no previously applied cast or splint was present in the radiograph, no complete elbow dislocation or displaced or comminuted fracture was present, and no metallic surgical hardware was present. Lateral radiographs were considered technically inadequate if the radiology report specifically stated that the lateral view was nondiagnostic or if anterior periarticular fat was not visible because of highly rotated elbow positioning. The lateral radiograph was saved as an anonymized portable network graphics file with window and level reset to "as acquired," and the original label of "effusion" or "no effusion" was recorded. A total of 901 images and labels were collected from 882 patients. Patient demographics are depicted in Table 1.

The images were randomly sorted and divided into a training set (657 images), validation set (115 images), and test set (127 images). All images from the 19 patients with bilateral elbow radiographs were sequestered into the training set.

To improve the accuracy of the labels (i.e., the originally reported diagnosis of effusion or no effusion), a subset of the total dataset was selected for further review by three fellowship-trained musculoskeletal radiologists with 3, 11, and 11 years of postfellowship clinical experience, respectively (denoted radiologists 1, 2, and 3). First, a preliminary review of the 657-image training set was conducted by a PGY 5 radiology resident, and 18 potentially mislabeled images were selected. These 18 images, the entire 115-image validation set, and the entire 127-image independent test set (a total of 262 images) were then relabeled by the expert reviewers. Judgments of effusion or no effusion were provided independently using a custom web interface that allowed reviewers to window, pan, and zoom the images. Reviewers were blinded to the originally reported diagnoses. One of the reviewers had previously interpreted one image during routine work, but more than 1 year had passed since the original interpretation. Im-

TABLE 1: Patient Demographics by Dataset

Characteristic	Training Set (*n* = 657)	Validation Set (*n* = 115)	Test Set (*n* = 129)	Total (*n* = 901)
Age				
Mean (SD) (y)	11.3 (5.2)	11.9 (4.8)	11.4 (5.4)	11.4 (5.1)
1–5 y	131 (19.9)	13 (11.3)	30 (23.3)	174 (19.3)
6–10 y	173 (26.3)	38 (33.0)	29 (22.5)	240 (26.6)
11–15 y	198 (30.1)	35 (30.4)	39 (30.2)	272 (30.2)
16–19 y	155 (23.6)	29 (25.2)	31 (24.0)	215 (23.9)
Sex				
Male	415 (63.2)	77 (67.0)	90 (69.8)	582 (64.6)
Female	242 (36.8)	38 (33.0)	39 (30.2)	319 (35.4)

Note—Except where noted otherwise, data are number (%) of patients. Not all percentages for a category total 100% because of rounding.

ages were presented at 512 × 512 pixel resolution. The majority judgment among the three expert reviewers was used as the new ground truth label for each image. To estimate the accuracy of the new labels, pairwise observer agreement between the three expert radiologist reviewers was measured [19]. Overall agreement, positive agreement, negative agreement, and the Cohen kappa statistic with 95% CIs were calculated using the fmsb package in the R programming language and statistical computing environment (version 3.4.3, R Foundation for Statistical Computing) [20].

Model Training and Selection

Models were trained and tested using Keras (version 2.1.3, François Chollet), a high-level neural network library, run on top of TensorFlow (version 1.5.0, Google Brain Team), an open-source deep learning framework, on a computer with a 3.7-GHz processor (Core i7–8700 K, Intel), 16-GB random access memory, and an 11-GB graphics card (GTX 1080 Ti, Nvidia).

Training DCNNs requires the choice of multiple variables that are set before training and affect how well the training process proceeds but are not part of the final trained model. These variables are called hyperparameters to distinguish them from the actual parameters in the final trained model. Examples of hyperparameters include network architecture, learning rate, and number of iterations through the training set (i.e., epochs) [21]. In this study, hyperparameters were systematically optimized by training 465 models with different hyperparameter settings and assessing performance on the validation set. For each of the 465 trials, hyperparameters were chosen randomly from a distribution of interest [22]. Initial learning rate and dropout rate were selected from uniform log-domains in early trials and were later selected from narrower uniform linear domains [21, 22]. Other continuous hyperparameters were selected from uniform linear domains. Integer hyperparameters were selected from uniform linear domains of integers. Boolean hyperparameters were selected from binary (true or false) domains. For each trained model produced during the hyperparameter search, the ROC curve was plotted by systematically varying the threshold (i.e., the cutoff point between 0 and 1 above which the model output is considered diagnostic of elbow effusion) and calculating true-positive rate and false-positive rate [23]. The ROC AUC was calculated using scikit-learn (version 0.19.1), an open-source machine learning library, and the model with the highest ROC AUC on the validation set was selected as the final model. Ninety-five percent CIs were calculated by use of the DeLong method using the pROC package in R [24–26].

Optimum operating threshold for calculation of sensitivity, specificity, and accuracy was chosen as the cutoff point that maximizes the Youden index (defined as sensitivity + specificity − 1) on the validation set [27].

Performance Measurement

The performance of the final trained model was measured by ROC AUC on the test set. Sensitivity, specificity, and accuracy were calculated using the optimum operating threshold from the validation set. Ninety-five percent CIs for sensitivity and specificity were calculated by the bootstrap method with 2000 stratified bootstrap replicates using the pROC package in R [25]. Ninety-five percent CIs for accuracy were calculated using the Clopper-Pearson method using the epiR package in R [8, 29].

To estimate model performance relative to nonradiologist physicians, a PGY 5 pediatric emer-

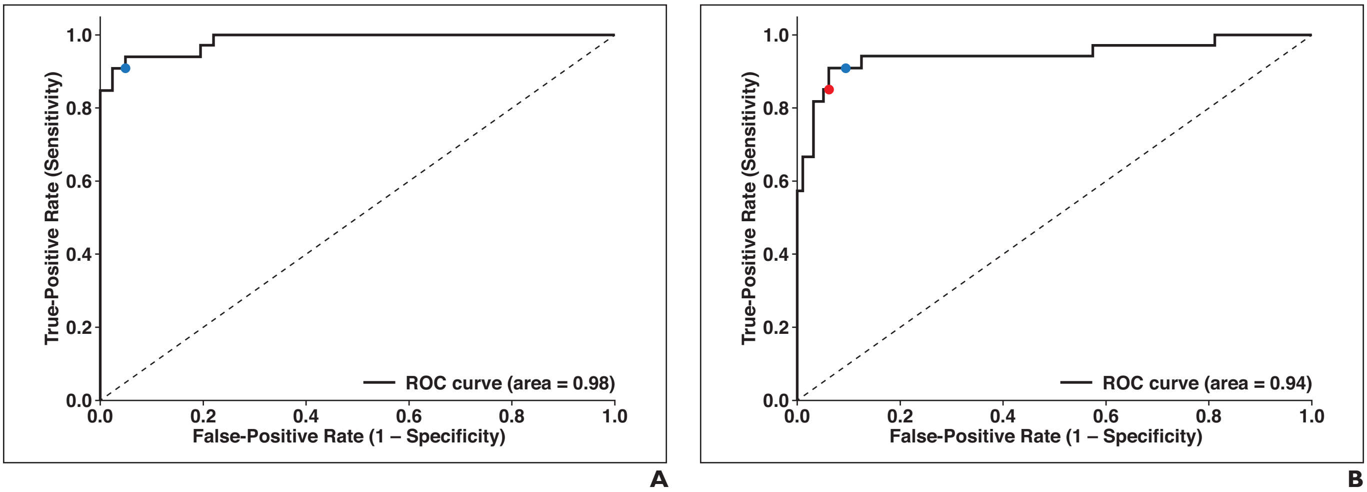

Fig. 2—ROC curves for two different datasets.
A and B, Graphs show ROC curves for validation set (**A**) and test set (**B**). Diagnostic performance using maximum Youden index for validation set as threshold is depicted as blue dot (**A** and **B**). Diagnostic performance of nonradiologist pediatric emergency medicine fellow on test set is depicted as red dot (**B**). Dashed line is line of no discrimination.

TABLE 2: Pairwise Observer Agreement Between the Three Subspecialty-Trained Musculoskeletal Radiologists

Comparison	Overall Agreement	Positive Agreement	Negative Agreement	Cohen κ Coefficient (95% CI)
Radiologist 1 vs radiologist 2	0.969	0.947	0.979	0.925 (0.874–0.976)
Radiologist 1 vs radiologist 3	0.947	0.908	0.962	0.870 (0.804–0.936)
Radiologist 2 vs radiologist 3	0.962	0.935	0.973	0.908 (0.852–0.964)

gency medicine fellow reviewed and labeled the test set using the same custom web interface as previously used by the expert reviewers. Sensitivity and specificity were calculated and compared with the final trained model. Ninety-five percent CIs for sensitivity, specificity, and accuracy were calculated by the Clopper-Pearson method.

A qualitative comparison of DCNN model and PGY 5 pediatric emergency medicine fellow performance on the test set was conducted by plotting the diagnostic operating point (x = false-positive rate, and y = true-positive rate) of the human observer on the ROC curve of the DCNN model, with a point below the curve interpreted as inferior performance, a point on the curve interpreted as essentially equivalent performance, and a point above the curve interpreted as superior performance.

Visualization

To gain intuition about which features in an image most strongly affect the DCNN prediction, saliency maps were constructed using the Keras-vis toolkit in [30].

Results

The judgments provided by the expert reviewers showed almost perfect agreement. Pairwise overall agreement, positive agreement, negative agreement, and Cohen kappa coefficients for the 262 reviewed images are presented in Table 2. When the majority rating was used as the ground truth, 21 of 262 images changed labels from the originally reported diagnosis: 11 of 18 training set images (three from no effusion to effusion and eight from effusion to no effusion), three of the 115 validation set images (all three from effusion to no effusion), and seven of the 129 test set images (four from no effusion to effusion and three from effusion to no effusion). The frequency of effusion in the relabeled datasets was 23.9% in the training set, 28.7% in the validation set, and 25.6% in the test set. The estimated lower limit of label accuracy for the relabeled images, predicated on the assumption that agreement implies accuracy, is 97.4–98.5% [19].

Multiple DCNNs were trained with varying hyperparameter settings. Some used newly initialized parameter weights by random sampling from a normalized gaussian distribution and others used preset parameter weights from previous training on an ImageNet dataset (i.e., transfer learning) [31, 32]. A glossary of relevant technical terms is presented in Table 3. After 465 trials, the best performing neural network model used the Dense Convolutional Network architecture with bottleneck layers and compression (DenseNet-BC), depth of 40 layers, growth rate of 12, and compression of 0.77 [33]. After random initialization of the model parameters, the model was trained for 544 epochs using the Adam optimization algorithm with fixed learning rate of 7.62×10^{-4}, minibatch size of 16, and dropout rate of 0.041 [32, 34, 35]. The training set images were resized to 224 × 224 pixel resolution and augmented using randomly applied horizontal flips, rotations of ± 19.8°, horizontal translations of ± 28.6%, vertical translations of ± 19.8%, zooming of ± 30.7%, and spatial shear transformations of ± 25.3%. Total training time was 1 hour, 1 minute, 24 seconds. Performance on the validation set (Fig. 2A) was an ROC AUC of 0.985 (95% CI, 0.966–1.000). Performance on the independent test set (Fig. 2B) was an ROC AUC of 0.943 (95% CI, 0.884–1.000). Results at the optimum operating threshold are presented in a confusion matrix in Figure 3A. Sensitivity was 0.909 (95% CI, 0.788–1.000; calculated using the bootstrap method, with 2000 stratified boot-

TABLE 3: Glossary of Technical Terms

Term	Explanation
Activation function	A nonlinear function applied between layers of an artificial neural network algorithm that allows the algorithm to capture highly complex nonlinear relationships in data.
Optimization algorithm	An iterative algorithm specifying how the artificial neural network parameters (see below) are updated during training. In a single step, the neural network uses the current parameters to classify the examples in a minibatch (see below), and then a measurement of the error is used to change (i.e., update) the parameters with the goal of achieving less error in the next step.
Deep convolutional neural network (DCNN)	A subtype of artificial neural network algorithms that uses a mathematic operation called convolution and contains many layers of computation.
Dropout	A method of preventing an artificial neural network from overfitting (i.e., becoming so accurate at correctly classifying cases in the training set that the algorithm does not generalize well to new cases outside the training set, suggesting that the algorithm will not perform well in real-world conditions).
Epoch	One complete iteration through the training set during the training of an algorithm. Training an artificial neural network requires many iterations through the training set data to become accurate.
Hyperparameters	Variables set before training and that affect how well the training process proceeds but are not part of the final trained neural network (see "Parameters" below).
Minibatch	A small subset of the training set used during the training of an algorithm, usually because of limited computational power and memory. Multiple minibatches of examples are presented to the algorithm during one epoch (see above).
Parameters	Numeric variables that are actually part of a final trained artificial neural network algorithm.

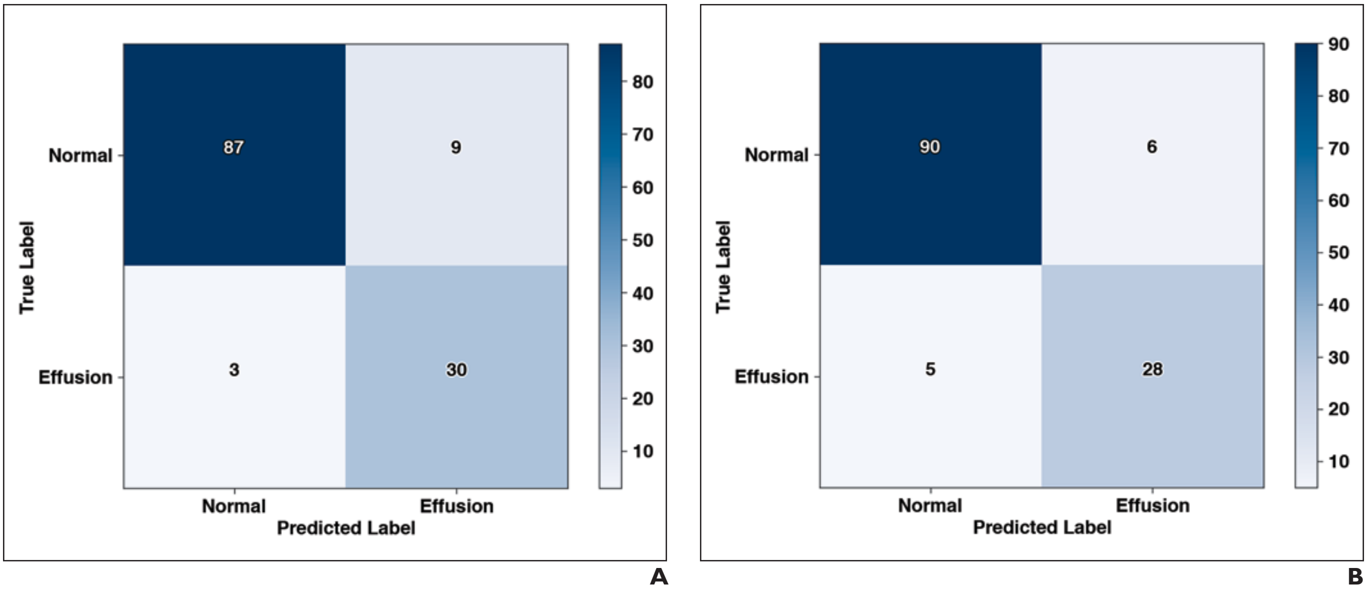

Fig. 3—Confusion matrices of diagnostic performance.
A and **B,** Graphs show diagnostic performance of trained deep convolutional neural network model (**A**) and nonradiologist pediatric emergency medicine fellow (**B**) on test set.

strap replicates), specificity was 0.906 (95% CI, 0.844–0.958; calculated using the bootstrap method, with 2000 stratified bootstrap replicates), and accuracy was 0.907 (95% CI, 0.843–0.951; calculated using the Clopper-Pearson method). Of three false-negative predictions, three were rated as effusion by all three expert reviewers. Of nine false-positive predictions, seven were rated as no effusion by all three expert reviewers, and two were rated as no effusion by two expert reviewers (i.e., at least one expert reviewer agreed with the positive prediction).

To estimate model performance relative to nonradiologist physicians, a PGY 5 pediatric emergency medicine fellow reviewed and labeled the test set. Results are presented in a confusion matrix in Figure 3B. Sensitivity was 0.848 (95% CI, 0.681–0.949), specificity was 0.938 (95% CI, 0.869–0.977), and accuracy was 0.915 (95% CI, 0.852–0.957), all calculated using the Clopper-Pearson method. The diagnostic operating point (*x* = false-positive rate, *y* = true-positive rate) corresponded to a point on the ROC curve of the DCCN model (Fig. 2B). Of five false-negative predictions, three were rated as effusion by all three expert reviewers, and two were rated as effusion by two expert reviewers (i.e., at least one expert viewer agreed with the negative prediction). Of six false-positive predictions, four were rated as no effusion by all three expert reviewers, and two were rated as no effusion by two expert reviewers (i.e., at least one expert reviewer agreed with the positive prediction).

In saliency maps of the test set images, most of the highly salient pixels cluster in the periarticular soft tissues, with many in the anterior and posterior fat pads (Figs. 4–7).

Discussion

Accurate diagnosis of traumatic elbow effusion is a clinically relevant binary classification task that immediately affects patient care. Even in the absence of identifiable fracture, patients with elbow joint effusion are treated presumptively with immobilization and follow-up radiography [36]. Evidence suggests that radiographic diagnosis of elbow joint effusion and acute elbow fracture is challenging for nonradiologists and is commonly missed [8]. In one study of radiographically apparent fractures missed by emergency medicine physicians, acute elbow fracture was among the most commonly missed fracture types (second only to rib fractures), making up 12% of the missed fractures, and with 68% showing traumatic elbow joint effusion that was overlooked [9].

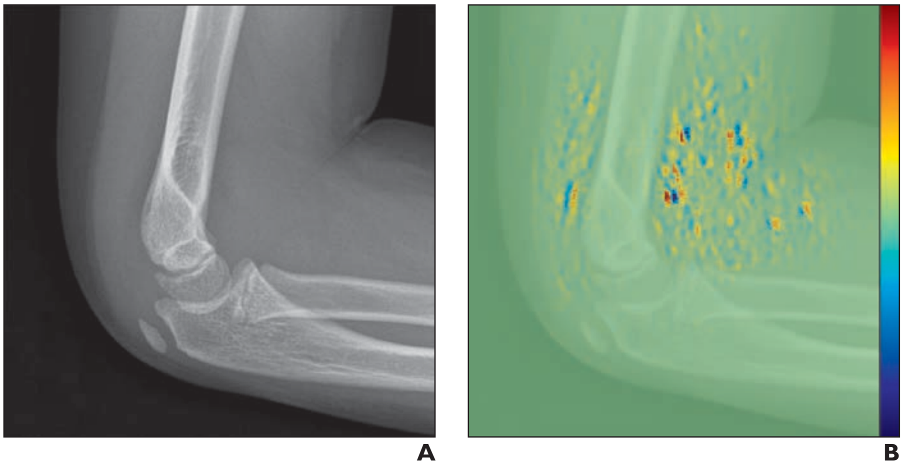

Fig. 4—9-year-old boy with acute right elbow effusion secondary to intraarticular fracture of radial head.
A, Lateral radiograph of right elbow shows elevation of anterior and posterior fat pads consistent with elbow effusion.
B, Saliency map shows high saliency (*red*) in periarticular soft tissues and in anterior and posterior fat pads and detects interfaces between fat pads and adjacent soft tissue as red-blue interfaces.

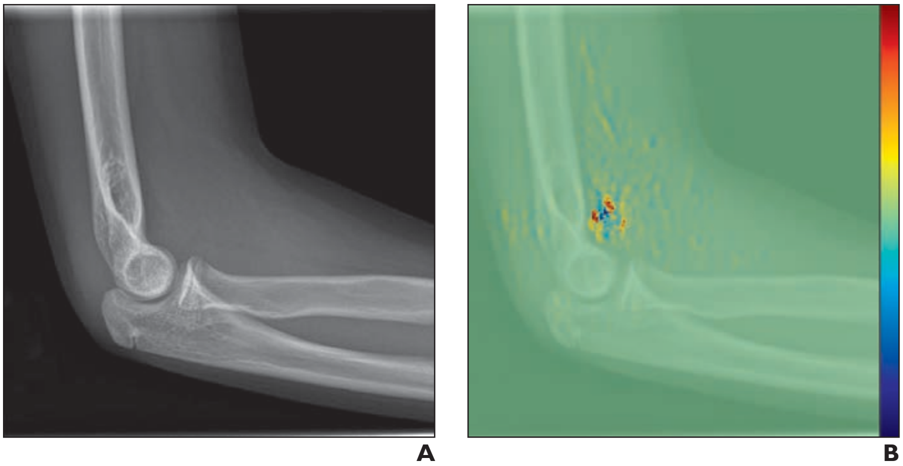

Fig. 5—11-year-old girl with acute right elbow effusion secondary to Salter-Harris type II avulsion fracture of medial epicondyle.
A, Lateral radiograph of right elbow shows elevation of anterior fat pad consistent with elbow effusion.
B, Saliency map shows high saliency (*red*) in anterior fat pad.

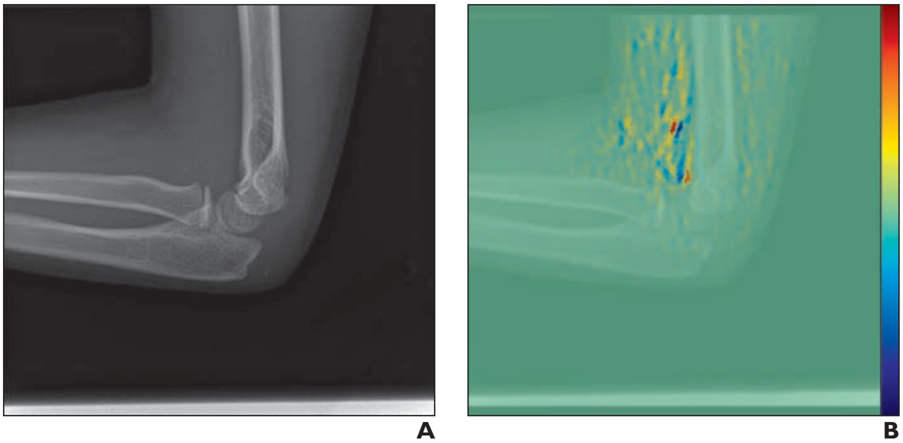

Fig. 6—9-year-old boy with no evidence of elbow effusion.
A, Lateral radiograph of left elbow shows thin normal anterior fat pad.
B, Saliency map shows high saliency (*red*) in thin anterior fat pad and detects interface between anterior fat pad and adjacent soft tissue as red-blue interface.

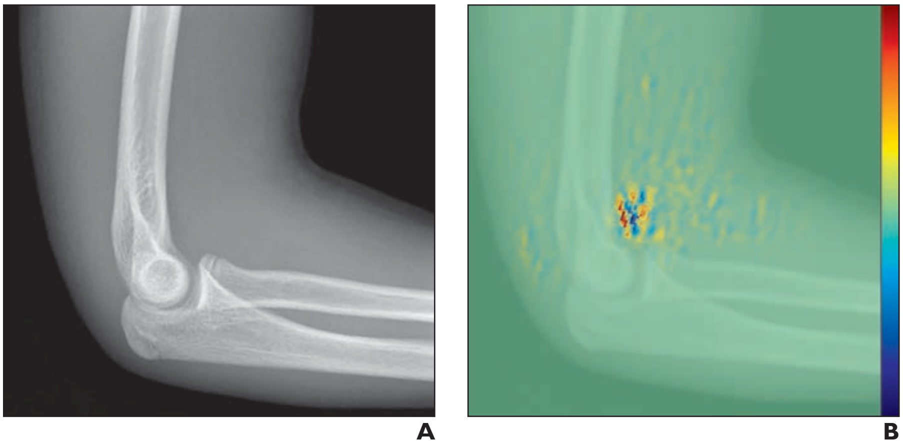

Fig. 7—14-year-old boy with no evidence of elbow effusion.
A, Lateral radiograph of right elbow shows thin normal anterior fat pad.
B, Saliency map shows high saliency (*red*) in thin anterior fat pad and detects interface between anterior fat pad and adjacent soft tissue as red-blue interface.

Accurate automated diagnosis of traumatic pediatric elbow joint effusion could therefore be useful as a decision support tool in the acute care setting, especially when a radiologist is not immediately available.

We present a deep learning model capable of human-level performance in the detection of traumatic pediatric elbow joint effusion. The sensitivity and specificity of a senior pediatric emergency medicine fellow on the test set corresponded to a point on the ROC curve of the deep learning model, suggesting equivalent performance, at least in a narrow operating range [23]. At the optimum operating threshold, the deep learning model showed slightly higher sensitivity and slightly lower specificity and accuracy, differences that were not statistically significant. A pediatric emergency medicine fellow may be expected to perform near the upper bound of diagnostic accuracy for nonradiologist acute care practitioners without specific pediatric training, including physicians, nurse practitioners, and physician assistants. Our DCNN model trained on a limited dataset could, therefore, already be of use to inexperienced nonradiologist clinicians. Furthermore, our deep learning model was trained to accommodate lateral elbow radiographs presented in their original state, without any user-dependent windowing or cropping, and could, therefore, also be useful in intelligent radiologist worklist management by flagging potentially positive cases for expedited reporting. Judging by the success of other deep learning models trained on far larger datasets, it is likely that further work with a larger dataset and higher computational power would result in an even better performing deep learning model.

Our model was trained on a dataset of limited size compared with most successful deep learning models, which usually require a large volume of data. For example, recently published deep learning models able to diagnose chest abnormalities on chest radiograph and hip fracture on hip radiograph used total datasets of 112,120 and 53,278 images, respectively [15, 17]. Two methods are commonly used to overcome the limitations of small dataset size. The first is data augmentation, where the training set is artificially enlarged via minor alterations in the images, such as horizontal flips, rotations, or translations. The second is by using a published high-performing network architecture with parameters that have already been trained on a different large dataset and refining a small subset of the parameters on a new smaller dataset. This latter approach has been termed "transfer learning," because many of the hierarchic features extracted by the refined model are learned from a different dataset and are simply transferred unaltered to the new model [31]. Transfer learning has been used successfully to achieve good performance in medically relevant tasks, even when training data are limited [16, 37]. Our model benefited from extensive data augmentation, but, notably, it did not use transfer learning and instead used a DenseNet-BC architecture that was initialized from scratch. Given the small size of our dataset, this is a surprising outcome that we hypothesize was made possible by the DenseNet architecture, a recently published state-of-the-art architecture that connects every layer to every subsequent layer, thus allowing highly efficient use of trained parameters with features extracted at early layers available as inputs for all subsequent layers [33].

A commonly discussed limitation of DCNNs is their black-box nature [38, 39]. Although deep learning models often show high performance on visual classification tasks, it is not always possible to understand exactly what the model is looking at. It is important to show that a high-performance model is actually detecting the correct region of the image and not overfitting to an associated finding, such as fracture in the case of traumatic elbow effusion, or to a superficially correlated finding, such as a marker placed by the radiography technologist to indicate maximum point of tenderness. One method of visualizing the areas of an image that are most important in determining the DCNN prediction is to construct a saliency map where each pixel in a given image is ranked according to how much it influences the prediction. High pixel saliency indicates that even a small change in pixel intensity will greatly affect the prediction [31]. Saliency maps of true-positive and true-negative examples from the test set (Figs. 4–7) show that our model predictions are most significantly affected by pixels in the periarticular soft tissues, with many of the most salient pixels in the anterior and posterior fat pads.

We acknowledge several limitations of our study. First, the small dataset size not only limited the training of our model but also limited the size of the test set, thus decreasing the statistical power of performance analysis. Although our results prove the feasibility of using a deep learning model for the detection of pediatric elbow effusion, further work expanding the dataset will allow more statistically rigorous assessment of performance. Second, the reference standard for diagnosis of elbow joint effusion was radiologist opinion, rather than a more definitive follow-up imaging modality such as MRI. However, advanced imaging is rarely performed to exclude acute pediatric elbow fracture, largely because of the added expense, the need for procedural sedation in young children, and increased length of stay for each patient encounter. Also, clinical and imaging follow-up is often performed at outside pediatric practices, thus limiting the opportunity for follow-up confirmation of the initial diagnosis. For these reasons, in clinical practice, radiologist opinion is essentially the standard for diagnosis of elbow effusion. Third, calculations of sensitivity, specificity, and accuracy required the choice of an explicit operating threshold. Our use of the maximum Youden index assumes that the harm of erroneously treating a patient without an effusion equals the harm of not treating a patient with an effusion [40]; however, we lack sufficient evidence on the harms of misdiagnosis to choose a more appropriate operating threshold. Anecdotally, physicians interpreting elbow radiographs tend to err on the side of increased sensitivity at the expense of specificity, but this may be due more to individual bias against missing a fracture than to a well-considered evaluation of harms.

In conclusion, this preliminary investigation shows that a DCNN model can achieve equal performance to a nonradiologist pediatric emergency medicine clinician in the diagnosis of traumatic pediatric elbow joint effusion, even when trained on a limited dataset. Further work developing a larger training dataset may result in even better performance.

References

1. Iyer RS, Thapa MM, Khanna PC, Chew FS. Pediatric bone imaging: imaging elbow trauma in children—a review of acute and chronic injuries. *AJR* 2012; 198:1053–1068
2. Goswami GK. The fat pad sign. *Radiology* 2002; 222:419–420
3. Norell HG. Roentgenologic visualization of the extracapsular fat: its importance in the diagnosis of traumatic injuries to the elbow. *Acta Radiol* 1954; 42:205–210
4. Bledsoe RC, Izenstark JL. Displacement of fat pads in diseases and injury of the elbow: a new radiographic sign. *Radiology* 1959; 73:717–724
5. Donnelly LF, Klostermeier TT, Klosterman LA. Traumatic elbow effusions in pediatric patients: are occult fractures the rule? *AJR* 1998; 171:243–245
6. Major NM, Crawford ST. Elbow effusions in trauma in adults and children: is there an occult fracture? *AJR* 2002; 178:413–418
7. Morewood DJ. Incidence of unsuspected fractures in traumatic effusions of the elbow joint. *Br Med J (Clin Res Ed)* 1987; 295:109–110
8. Wu J, Perron AD, Miller MD, Powell SM, Brady WJ. Orthopedic pitfalls in the ED: pediatric supracondylar humerus fractures. *Am J Emerg Med* 2002; 20:544–550
9. Freed HA, Shields NN. Most frequently overlooked radiographically apparent fractures in a teaching hospital emergency department. *Ann Emerg Med* 1984; 13:900–904
10. Goodfellow I, Bengio Y, Courville A. *Deep learning*. Cambridge, MA: MIT Press, 2016
11. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Adv Neural Inf Process Syst* 2012; 1:1097–1105
12. Chartrand G, Cheng PM, Vorontsov E, et al. Deep learning: a primer for radiologists. *RadioGraphics* 2017; 37:2113–2131
13. Szegedy C, Liu W, Jia Y, et al. Going deeper with convolutions. IEEE website. ieeexplore.ieee.org/document/7298594/. Published 2015. Accessed August 8, 2018
14. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv website. arxiv.org/abs/1409.1556. Published 2014. Accessed August 8, 2018
15. Gale W, Oakden-Rayner L, Carneiro G, Bradley AP, Palmer LJ. Detecting hip fractures with radiologist-level performance using deep neural networks. arXiv website. arxiv.org/abs/1711.06504. Published 2017. Accessed August 8, 2018
16. Lakhani P, Sundaram B. Deep learning at chest radiography: automated classification of pulmonary tuberculosis by using convolutional neural networks. *Radiology* 2017; 284:574–582
17. Rajpurkar P, Irvin J, Zhu K, et al. CheXNet: radiologist-level pneumonia detection on chest X-Rays with deep learning. arXiv website. arxiv.org/abs/1711.05225. Published 2017. Accessed August 9, 2018
18. Rajpurkar P, Irvin J, Bagul A, et al. MURA dataset: towards radiologist-level abnormality detection in musculoskeletal radiographs. ResearchGate website. www.researchgate.net/publication/321936621_MURA_Dataset_Towards_Radiologist-Level_Abnormality_Detection_in_Musculoskeletal_Radiographs. Published 2017. Accessed August 9, 2018
19. Kundel HL, Polansky M. Measurement of observer agreement. *Radiology* 2003; 228:303–308
20. Nakazawa M. fmsb: Functions for medical statistics book with some demographic data website. CRAN.R-project.org/package=fmsb. R package

version 0.6.1. Published 2017. Accessed March 5, 2018

21. Bengio Y. Practical recommendations for gradient-based training of deep architectures. In: Montavon G, Orr GB, Müller K-R, eds. *Neural networks: tricks of the trade*, 2nd ed. Berlin, Germany: Springer Berlin Heidelberg, 2012:437–478
22. Bergstra J, Bengio Y. Random search for hyperparameter optimization. *J Mach Learn Res* 2012; 13:281–305
23. Obuchowski NA. Receiver operating characteristic curves and their use in radiology. *Radiology* 2003; 229:3–8
24. DeLong ER, DeLong DM, Clarke-Pearson DL. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 1988; 44:837–845
25. Robin X, Turck N, Hainard A, et al. pROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 2011; 12:77
26. Pedregosa F, Varoquaux G, Gramfort A, et al. Scikit-learn: machine learning in python. *J Mach Learn Res* 2011; 12:2825–2830
27. Youden WJ. Index for rating diagnostic tests. *Cancer* 1950; 3:32–35
28. Newcombe RG, Altman DG. Proportions and their differences. In: Altman DG, Machin D, Bryant T, Gardner MJ, eds. *Statistics with confidence: confidence intervals and statistical guidelines*. Hoboken, NJ: Wiley, 2013:45–56
29. Stevenson M et al. epiR: Tools for the analysis of epidemiological data website. CRAN.R-project.org/package=epiR. R package version 0.9-93. Published 2017. Accessed March 5, 2018
30. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. arXiv website. arxiv.org/abs/1312.6034. Published 2013. Accessed August 9, 2018
21. Yosinski J, Clune J, Bengio Y, Lipson H. How transferable are features in deep neural networks? In: Xing EP, Jebara T, eds. *Proceedings of the 31st International Conference on Machine Learning*. Beijing, China: International Machine Learning Society, 2014:3320–3328
32. He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on ImageNet classification. arXiv website. arxiv.org/abs/1502.01852. Published 2015. Accessed August 9, 2018
33. Huang G, Liu Z, Weinberger KQ, van der Maaten L. Densely connected convolutional networks. arXiv website. arxiv.org/abs/1608.06993. Published 2017. Accessed August 9, 2018
34. Kingma DP, Ba J. Adam: a method for stochastic optimization. Cornell University Library website. arxiv.org/abs/1412.6980. Published 2014. Accessed August 9, 2018
35. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R. Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 2014; 15:1929–1958
36. Skaggs D, Pershad J. Pediatric elbow trauma. *Pediatr Emerg Care* 1997; 13:425–434
37. Cheng PM, Tejura TK, Tran KN, Whang G. Detection of high-grade small bowel obstruction on conventional radiography with convolutional neural networks. *Abdom Radiol (NY)* 2018; 43:1120–1127
38. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. In: Fleet D, ed. *ECCV 2014*. Geneva, Switzerland: Springer International Publishing, 2014:818–833
39. Beam AL, Kohane IS. Big data and machine learning in health care. *JAMA* 2018; 319:1317–1318
40. Habibzadeh F, Habibzadeh P, Yadollahie M. On determining the most appropriate test cut-off value: the case of tests with continuous results. *Biochem Med (Zagreb)* 2016; 26:297–307